

**APPLICATION FOR UNITED STATES LETTERS PATENT**

**FOR**

**SYSTEM AND METHOD FOR MEDIAN FUSION OF DEPTH MAPS**

**by**

**DAVID NISTER**

09891238.062704  
FO/290" 8E2T6860

BURNS, DOANE, SWECKER & MATHIS, L.L.P.  
POST OFFICE BOX 1404  
ALEXANDRIA, VIRGINIA 22313-1404  
(703) 836-6620  
Attorney's Docket Number 040000-755

### CROSS-REFERENCE TO RELATED APPLICATION

[0001] This application is based upon and claims priority from United States provisional application No. 60/214,791, filed June 28, 2000, the contents being incorporated herein by reference.

## BACKGROUND OF THE INVENTION

## Field of the Invention

**[0002]** The present invention relates generally to systems for creating graphical models from video sequences, and more particularly, to a system for constructing a refined graphical representation of an image from a plurality of calibrated camera views.

## Background Information

[0003] Computer-aided imagery is the process of rendering new two-dimension and three-dimension images on a terminal screen or graphical user interface from two or more digitized two-dimension images with the assistance of the processing and data handling capabilities of a computer. Constructing a three-dimension (hereinafter "3D") model from two-dimension (hereinafter "2D") images is utilized, for example, in computer-aided design (hereinafter "CAD"), 3D teleshopping, and virtual reality systems, in which the goal of the processing is a graphical 3D model of an object or a scene that was originally represented only by a finite number of 2D images. Under this application of computer graphics or computer vision, the 2D images from which the 3D model is constructed represent views of the object or scene as perceived from different views or locations around the object or scene. The images are obtained either from multiple cameras positioned around the object or scene or from a single camera in motion around the object, recording pictures or a video stream of images of the object. The information in the 2D images is combined and contrasted to produce a composite, computer-based graphical 3D model. While recent advances in computer processing power and data-handling capability have improved computerized 3D modeling, these graphical 3D construction systems remain characterized by demands for heavy computer processing power, large data storage requirements, and long processing

# SECRET

times. Furthermore, volumetric representations of space, such as a graphical 3D model, are not easily amenable to dynamic modification, such as combining the 3D model with a second 3D model or perceiving the space from a new view or center of projection.

[0004] Typically the construction of a 3D image from multiple views or camera locations first requires camera calibration for the images produced by the cameras to be properly combined to render a reasonable 3D reconstruction of the object or scene represented by the images. Calibration of a camera or a camera location is the process of obtaining or calculating camera parameters at each location or view from which the images are gathered, with the parameters including such information as camera focal length, viewing angle, pose, and orientation. If the calibration information is not readily available, a number of calibration algorithms are available to calculate the calibration information. Alternatively, if calibration information is lacking, some graphical reconstruction methods estimate the calibration of camera positions as the camera or view is moved from one location to another. However, calibration estimation inserts an additional variable in the 3D graphical model rendering process that can cause inaccuracies in the output graphics. Furthermore, calibration of the camera views necessarily requires prior knowledge of the camera movement and/or orientation, which limits the views or images that are available to construct the 3D model by extrapolating the calibrated views to a new location.

[0005] One current method of reconstructing a graphical 3D model of an object from multiple views is by using pairs of views of the object at a time in a process known as stereo mapping, in which a correspondence between the two views is computed to produce a composite image of the object. However, shape information recovered from only two views of an object is neither complete nor very accurate, so it is often necessary to incorporate images from additional views to refine the shape of the 3D model. Additionally, the shape of the stereo mapped 3D model is often manipulated in some graphical systems by the weighting, warping, and/or blending of one or more of the images to adjust for known or perceived inaccuracies in the image or calibration data. However, such manipulation is a manual process, which not only limits the automated computation of composite graphical images but also risks introducing errors as the appropriate level of weighting, warping, and/or

09691233-062701

blending is estimated.

[0006] Recently, graphical images in the form of depth maps have been applied to stereo mapping to render new 2D views and 3D models of objects and scenes. A depth map is a two-dimension array of values for mathematically representing a surface in space, where the rows and columns of the array correspond to the x and y location information of the surface; and the array elements are depth or distance readings to the surface from a given point or camera location. A depth map can be viewed as a grey scale image of an object, with the depth information replacing the intensity information, or pixels, at each point on the surface of the object. Accordingly, surface points are also referred to as pixels within the technology of 3D graphical construction, and the two terms will be used interchangeably within this disclosure.

[0007] A graphical representation of an object can be estimated by a depth map under stereo mapping, using a pair of views at a time. Stereo depth mapping typically compares sections of the two depth maps at a time, attempting to find a match between the sections so as to find common depth values for pixels in the two maps. However, since the estimated depth maps invariably contain errors, there is no guarantee that the maps will be consistent with each other and will match where they should. While an abundance of data may be advantageous to minimize the effect of a single piece of bad or erroneous data, the same principle does not apply to depth maps where any number of depth maps may contain errors because of improper calibration, incorrect weighting, or speculations regarding the value of the particular view, with any errors in the depth maps being projected into the final composite graphical product. Furthermore, conventional practices of stereo mapping with depth maps stop the refinement process at the estimation of a single depth map.

[0008] The preferred embodiments of the present invention overcome the problems associated with existing systems for fusing a number of depth maps into a consistent and accurate representation of a three dimensional object or scene.

## SUMMARY OF THE INVENTION

[0009] The present invention is directed toward a system and method for computing a refined depth map for a spatial representation from a plurality of estimated depth maps. The system utilizes information in multiple depth maps of a represented image to refine an optimum depth map, selecting the median depth value for each pixel in each depth map rendered into a new view. The system determines the median depth map value for each pixel by selecting the rendered pixel depth value that is the median value among all the respective pixels from the rendered depth maps, which are fused one at a time with the new view of the object or scene. The resulting median fused depth map can be used to more accurately represent a graphical or video image as perceived from the new view.

[0010] In accordance with one aspect of the present invention, a system and method is directed toward computing a fused depth map of a view of an object, including obtaining camera calibration data for a new view of an object; rendering each of a plurality of known depth maps of the object into the new view of the object based on the camera calibration of the new view; and computing the depth map elements of a fused depth map of the new view of the object, wherein each said depth map element corresponds to a pixel in the new view of the object and wherein each said depth map element comprises a median value of a set of rendered depth map elements, corresponding to the pixel, each from one of the plurality of rendered depth maps.

[0011] In accordance with another aspect of the invention, a system and method is directed toward constructing a graphical representation of an object, as perceived from a desired view, from a plurality of calibrated views, including obtaining camera calibration data for a desired view of the object; computing a plurality of depth maps, each depth map respectively corresponding to each of a plurality of calibrated views of the object; for each of the plurality of computed depth maps, rendering a selected, computed depth map into the desired view; for each pixel in the desired view, computing a median fused depth map value at said pixel, wherein as many rendered depth maps occlude said pixel as pass it by; and projecting the graphical representation of the object as perceived from the desired view, utilizing the computed median fused depth map values.

## BRIEF DESCRIPTION OF THE DRAWINGS

[0012] These and other objects and advantages of the present invention will become more apparent and more readily appreciated to those skilled in the art upon reading the following detailed description of the preferred embodiments, taken in conjunction with the accompanying drawings, wherein like reference numerals have been used to designate like elements, and wherein:

Figure 1 shows a depth map of view  $i$ , wherein a point,  $x$ , is located on the target surface in view  $a$ ;

Figure 2 shows a depth map of view  $i$ , wherein a point,  $x$ , is occluded in view  $a$ ;

Figure 3 shows a depth map of view  $i$ , wherein a point,  $x$ , is passed in view  $a$ ;

Figure 4 shows a block flow chart of an exemplary method for constructing a median fused depth map of a view; and

Figure 5 illustrates the devices and communication links of an exemplary median depth map fusion system.

## DETAILED DESCRIPTION OF THE PREFERRED EMBODIMENTS

[0013] In the following description, for purposes of explanation and not limitation, specific details are set forth in order to provide a thorough understanding of the present invention. However, it will be apparent to one skilled in the art that the present invention may be practiced in other embodiments that depart from these specific details. In other instances, detailed descriptions of well-known methods, devices, and circuits are omitted so as not to obscure the description of the present invention.

[0014] While the present invention can be utilized for determining fused depth maps of virtually any object or scene and for constructing 3D images of such objects and scenes, the discussion below will refer to all such 3D items as "objects" to simplify the explanation of the embodiments of the invention. Referring now initially to Figure 1, there is shown a portion of an object as perceived from view  $i$ . The line segment,  $d_i$  106 represents the view of the object as seen from view  $i$ , in the image plane formed by lines  $L_1$  102 and  $L_2$  104,

emanating from a center of projection (hereinafter "COP") of view  $i$  at  $P_i$  100. The view  $i$  represents one of a plurality of known views of the object, for which camera calibrations are known or can be calculated. View  $a$  represents a new, desired orientation from which to view the object and for which a preferred embodiment of the present inventive system computes a median fused depth map as an optimized depth map for the view  $a$ , given the known views and the camera calibration at  $P_a$  108, the COP for view  $a$ . While the invention is presently envisioned to be implemented on a personal or laptop computer, the method and apparatus of the present median depth map fusion system can function equally well on any number of computers or processors without detracting from the inventive features of the system.

[0015] In a preferred embodiment of the present invention for computing a median fused depth map of an object, a set of  $n$  calibrated views,  $\{v_i\}_{i=1}^n$ , of the object is first

obtained by any one of several well known methods for capturing views of an object. While the invention is operable given a minimum of two views of the object, the resultant fused depth map will be a more accurate graphical rendition of the object if a greater number of views are utilized in the computation of the fused depth map.

[0016] For each view  $i$  (represented as  $v_i$ ) of the object, there is an associated depth function  $d_i$  defined on the domain of the image of the object. For each pixel  $x$  110 in view  $i$ , the depth function has a depth map value of  $d_i(x)$ . The depth map value  $d_i(x)$  is an estimate of the distance from the center of projection for view  $i$  to the closest point or pixel  $x$  along the backprojected ray corresponding to the image point  $x$  in view  $i$  of the object. Referring to Figure 1, the depth value  $d_i(x)$  for pixel  $x$  110 in view  $i$  is the distance from COP  $P_i$  100 to pixel  $x$  110.

[0017] Referring now to Figures 1, 2, and 3, any pixel  $x$  110, 200, or 300 in the image domain of the object is either on the surface  $d_i$  as in Figure 1, is located beyond the surface from  $P_i$ , as shown in Figure 2, or is located "in front of" the surface, relative to  $P_i$ , as shown in Figure 3. Points located beyond  $d_i$  from  $P_i$ , as in Figure 2, are said to be "occluded" by depth map  $d_i$  in view  $a$  because  $P_a$  cannot "see" point  $x$  200 because of the surface  $d_i$ . Points located in front of  $d_i$ , as shown in Figure 3, are said to be "passed" by depth map  $d_i$  in view  $i$ . When

viewed as a continuous function, the depth map  $d_i$  defines a surface in space as perceived from view  $i$ , and as graphically represented by element  $d_i$  106 in the planar views of Figures 1, 2, and 3.

[0018] The present invention utilizes a variation of the mathematical principle of median values to compute an optimum depth map,  $f_a$ , of a new, desired view  $a$  from a plurality of estimated depth maps  $\{d_i\}_{i=1}^n$  corresponding to known views  $\{v_i\}_{i=1}^n$  of the object.

This optimum depth map, as computed by embodiments of the present invention, is termed a "median fused depth map" because of the process described below of fusing a plurality of depth maps such that a median depth value is determined for each pixel  $x_a$  110 on the surface of the object in the new view  $a$ . The median fused depth map value of a pixel  $x_a$  in new view  $a$  is defined to be "stable for view  $a$ " when, in the set of known depth maps  $\{d_i\}_{i=1}^n$  for the

object, there are at least as many depth maps that occlude  $x_a$  200 in view  $a$  as there are depth maps that pass  $x_a$  300 in view  $a$ . Therefore, the process of the present invention is determining an optimum depth value for each pixel  $x_a$  110 in the new view  $a$  such that there are at least as many known depth maps  $d_i$  that occlude  $x_a$  200 in view  $a$  as there are depth maps  $d_i$  that pass  $x_a$  300 in view  $a$ . The optimum, or median fused depth value,  $f_a(x_a)$ , for each pixel  $x_a$  110 in view  $a$  is the infimum, or greatest lower value, of all depth values  $d_i(x)$  such that all pixels  $\{x_a\}_{a=1}^s$  110 are stable for view  $a$ .

[0019] Prior to explaining the computation of a median fused depth map by a preferred embodiment of the present invention, definitions of several working elements of the system will be presented. The median fused depth map of view  $a$  as computed by the system is represented as  $f_a$ , with the median fused depth map value for a pixel  $x$  in view  $a$  is represented as  $f_a(x)$ . Referring to the flow chart of Figure 4, one of the first steps of the median depth map fusion process, as discussed below regarding step 400, is to establish an



estimated lowest value for  $f_a$ . Three two-dimension arrays are utilized to hold working values of minimum rendered depth map values, occlusion indicators, and fused depth values during the processing of all the depth maps  $\{d_i\}_{i=1}^n$ . A minimum depth map value matrix  $o(x_a)$  is

utilized to hold the minimum depth map values for the pixels,  $x_a$ , of view  $a$ . As each new depth map,  $d_i$ ,  $i=1, \dots, n$ , is rendered into view  $a$ , the smallest of the depth map values for pixel  $x_a$  of depth map  $d_i$  as rendered into view  $a$  is stored as  $o(x_a)$ . As each subsequent depth map,  $d_i$ ,  $i=1, \dots, n$ , is rendered into view  $a$ , the values or elements of the minimum depth map value matrix  $o(x_a)$  are reset. A support matrix,  $s(x_a)$ , defined on the image domain of view  $a$ , holds the number of "votes" regarding whether any target pixel  $x_a$  in view  $a$  is passed by or occluded by the various depth maps  $\{d_i\}_{i=1}^n$ .  $s(x_a)$  is termed a support matrix in that a

positive value in  $s(x_a)$  represents support for a higher depth map value at pixel  $x_a$ . That is,  $x_a$  is being passed by the depth map  $d_i$  in view  $a$ .  $s(x_a)$  is set to zero at step 404 at the beginning of each iteration of the processing of the set of depth maps  $\{d_i\}_{i=1}^n$ . During the computation

of the median fused depth map, and for each pixel  $x_a$  represented in view  $a$ , the support matrix value at  $s(x_a)$  is subtracted by 1 if the depth map  $d_i$  rendered into view  $a$  occludes point  $x_a$ . If the rendered depth map passes by  $x_a$  in view  $a$ , the corresponding support matrix value  $s(x_a)$  is incremented by one, as discussed below. A next candidate depth map  $n(x_a)$  is the repository of the computed, median fused depth value for each pixel  $x_a$  as the present invention processes through the set of depth maps  $\{d_i\}_{i=1}^n$  and represents the next possible

candidate for the depth map at each pixel  $x_a$ . The next candidate depth map values are initially set to be equal to  $\infty$  at step 406.

[0020] The target median fused depth value,  $f_a(x_a)$ , for each pixel  $x_a$  in view  $a$  is computed as follows. Given two or more depth maps  $\{d_i\}_{i=1}^n$  of respective views  $\{v_i\}_{i=1}^n$  and





map has been processed, the depth map values for the median fused depth map are set at step 426 by evaluating the value of the support matrix for each pixel  $x_a$  110, 200, and 300 in the domain of view  $a$  as follows:

For all pixels  $x_a$  in view  $a$ , if  $s(x_a) > 0$  then set  $f_a(x_a) = n(x_a)$  [4]

[0027] If the value of the support matrix at pixel  $x_a$  is greater than zero, the net votes from the processing of the known depth maps rendered into view  $a$  indicate a larger depth value than the minimum depth value, the larger value having been computed into the next candidate depth map  $n(x_a)$  for pixel  $x_a$ . If the value of the support matrix is less than or equal to zero, the net processing of the rendered depth maps has concluded a smaller depth map value for pixel  $x_a$  is called for; and the initialized minimum value in  $f_a(x_a)$  as computed at step 404 is retained. With an updated, refined set of values for the fused depth map  $f_a$ , processing of all the depth maps is reinitiated at step 402 until the full set of depth maps has been rendered into the new view  $a$   $n-1$  times, guaranteeing convergence with the processing of steps 408 - 426  $n^2$  times.

[0028] Referring now to Figure 5, there are illustrated the devices and communication links of an exemplary median depth map fusion system in accordance with the present invention. The components of Figure 5 are intended to be exemplary rather than limiting regarding the devices and communication pathways that can be utilized in the present inventive system. The building 500 represents the object or scene that can be the subject image of the views of the present invention, and cameras 502 - 510 represent the plurality of views or locations from which the object 500 can be perceived and for which depth maps are available. View 512 represents the new view for which the present inventive system can compute a median fused depth map, the calibration data associated with view 512 being known. The respective images captured from views 502 - 510 are transmitted or downloaded to a processor 514 for ultimate storage on database 516. The calibration information for views 502 - 512 can correspondingly be transmitted to processor 514, can be calculated by processor 518, and/or can be retrieved from database storage 516. The depth maps for views

502 - 510 are calculated or estimated by one of a plurality of known depth map calculation techniques and are stored on database 516. Alternatively, the processing functions of processor 514 and the storage functions of database 516 can be combined in a single device, such as a personal computer, as represented by processor 518; or can be distributed across multiple processors 514 and 518 and databases 516.

[0029] The actual computation of the median fused depth map is performed by processor 518, with data drawn from and stored onto database 516. In an alternative aspect of the invention, the storage provided by database 516 can be integrated into the processor 518. The various functional aspects of the present invention and the corresponding apparatus portions of the system for computing median fused depth maps, such as the initialization devices, selection devices, merging devices, and processors, can reside in a single processor 518 or can be distributed across a plurality of processors 518 and storage devices 516.

[0030] Once the median fused depth map is computed by processor 518 and stored on database 516, it can be accessed by any number of authorized users operating processors 522. These users can display a 2D or 3D representation of the median fused depth map on the screen or graphical user interface of the processor 522 and/or can print the same on the printer 524.

[0031] The communication and data links shown in Figure 5 connecting the various apparatus components of the system can be either wireless or hard-wired links or any combination thereof. Similarly, the inventive system as illustrated in Figure 5 can be implemented as a stand alone system or can be implemented across any known network or combination of networks, such as a local area network, a wide area network, an intranet, or the Internet, or any combination thereof, without detracting from the inventive features of the present median depth map fusion system.

[0032] Although preferred embodiments of the present invention have been shown and described, it will be appreciated by those skilled in the art that changes may be made in these embodiments without departing from the principle and spirit of the invention, the scope of which is defined in the appended claims and their equivalents.